Reproductive Toxicology 62 (2016) 92–99

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

Aggregate entropy scoring for quantifying activity across endpoints with irregular correlation structure

Guozhu Zhang [a], Skylar Marvel [a], Lisa Truong [c], Robert L. Tanguay [c], David M. Reif [a,b,*]

[a] Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA
[b] Department of Biological Sciences, Center for Human Health and the Environment, North Carolina State University, Raleigh, NC, USA
[c] Department of Environmental and Molecular Toxicology, Sinnhuber Aquatic Research Laboratory, Oregon State University, Corvallis, OR, USA

ARTICLE INFO

Article history:
Received 30 December 2015
Received in revised form 23 March 2016
Accepted 15 April 2016
Available online 27 April 2016

Keywords:
Developmental neurotoxicology
Chemical biology
Morphology
Zebrafish
High throughput screening
ToxCast
Multiplexed assays

ABSTRACT

Robust computational approaches are needed to characterize systems-level responses to chemical perturbations in environmental and clinical toxicology applications. Appropriate characterization of response presents a methodological challenge when dealing with diverse phenotypic endpoints measured using *in vivo* systems. In this article, we propose an information-theoretic method named Aggregate Entropy (AggE) and apply it to scoring multiplexed, phenotypic endpoints measured in developing zebrafish (*Danio rerio*) across a broad concentration-response profile for a diverse set of 1060 chemicals. AggE accurately identified chemicals with significant morphological effects, including single-endpoint effects and multi-endpoint responses that would have been missed by univariate methods, while avoiding putative false-positives that confound traditional methods due to irregular correlation structure. By testing AggE in a variety of high-dimensional real and simulated datasets, we have characterized its performance and suggested implementation parameters that can guide its application across a wide range of experimental scenarios.

1. Introduction

Biological responses in whole animals are the product of coordinated actions (or, in the case of toxic responses, dysregulation) on a systemic level. Accordingly, experimental inquiries into basic biological processes should record multiple phenotypic outcomes when assessing perturbations, from clinical interventions such as drug treatments to environmental stressors such as manufactured chemicals. Innovations in multiplexed endpoint measurement technology and exploratory omics platforms have enabled theoretically comprehensive experiments to be conducted [1]. However, these new, multi-endpoint data present challenges with respect to recapitulating the relevant biological processes: (1) The correlation structure across endpoints is irregular; (2) Individual subjects/samples vary in endpoint presentation; (3) Endpoint measurement methods are imperfect; (4) Experimental questions may depend on subsets and/or recombinations of endpoints. Therefore, analysis methods are needed that can address these challenges while allowing for either focused, *a priori* analysis or data-wide, empirical analysis.

One such area where comprehensive analysis of systemic response is needed is environmental and clinical toxicology, where adverse responses may manifest anywhere from specific abnormalities to collections of several endpoints that count as toxicity in the aggregate. While there is an ever-increasing number of chemicals in commerce and the environment, comprehensive toxicological knowledge is lacking for all but a handful of compounds—mostly pharmaceuticals that have progressed to expensive, late-stage clinical trials. Traditional animal testing is very expensive in terms of labor, time, and money, so high-throughput screening (HTS) is being developed in order to more efficiently assess chemical biocompatibility [2]. Experimental HTS includes both *in vitro* assays that probe molecular action and *in vivo* assays that screen for a

Abbreviations: MORT, mortality; YSE, yolk sac edema; AXIS, body axis; EYE, eye; SNOU, snout; JAW, jaw; OTIC, otic vesicle; PE, pericardial edema; BRAI, brain; SOMI, somite; PFIN, pectoral fin; CFIN, caudal fin; PIG, pigment; CIRC, circulation; TRUN, truncated body; SWIM, swim bladder; NC, notochord & bent tail; TR, touch response; NOAE, no observed adverse effect; AggE, Aggregate Entropy; PP, positive-positive; NN, negative-negative; NP, negative-positive; PN, positive-negative; Any End, any endpoint; ROC, receiver operating characteristic; hpf, hours post fertilization; AOP, adverse outcome pathway; HTS, high throughput screening; MI, mutual information; SE, super endpoint.

* Corresponding author at: Bioinformatics Research Center, Center for Human Health and the Environment, Department of Biological Sciences, North Carolina State University, Raleigh, NC, Box 7566, 1 Lampe Drive, Raleigh NC 27695, USA.
E-mail addresses: gzhang6@ncsu.edu (G. Zhang), swmarvel@ncsu.edu (S. Marvel), lisa.truong@oregonstate.edu (L. Truong), robert.tanguay@oregonstate.edu (R.L. Tanguay), dmreif@ncsu.edu (D.M. Reif).

http://dx.doi.org/10.1016/j.reprotox.2016.04.012

variety of phenotypic endpoints that cover fundamental developmental, structural, and neurological pathways [3–5].

These HTS *in vivo* assays provide an ideal workbench for the development and testing of analysis methods for multiple endpoints, in that the data can be generated on a scale that permits evaluation of an analysis method's ability to address the four challenges presented above. In particular, experimental methods for the zebrafish (*Danio rerio*), a model organism whose fundamental developmental processes are shared across vertebrates and that has high genetic similarity to humans, have exploded in recent years [6,7]. Several endpoints, ranging from specific structural features through outright mortality, have been measured, with a trend toward higher-order assessment of multiple endpoints during embryonic development [8].

Here, we developed an information theory-based method named Aggregate Entropy ("AggE") to consolidate information into classes across endpoints, then tested this method using both simulated and empirical zebrafish data. We characterized the relationship amongst endpoints to identify the biological processes underlying overall developmental assessments; used simulated data to further validate our method across a range of sample sizes; characterized the irregular correlation structure across endpoints using mutual information and normalized information distance; and used this information to reduce noise by collapsing endpoints with similar phenotypic response patterns. Finally, we parameterized AggE distributions to allow for application to new datasets of varying dimensions from multi-endpoint experiments in any model system.

2. Materials and methods

2.1. Empirical data

The empirical data were collected as described in Truong et al. [5] and Noyes et al. [9]. Fig. 1 shows the experimental design and data structure. The data include 1060 unique ToxCast chemicals tested at six concentrations for each chemical (0 μM, 0.0064 μM, 0.064 μM, 0.64 μM, 6.4 μM and 64 μM). There were n = 32 replicates (individual embryo wells) at each concentration. At 120 h post fertilization (hpf), 18 distinct developmental endpoints were evaluated. The data were recorded as binary incidences.

As in Fig. 1(B) and (C), we constructed 19 different biological states, including 18 developmental endpoints plus one NOAE (No Observed Adverse Effect) state. Thus, for each embryo per chemical-per concentration, data were shown as 0 and 1 for 18 binary endpoints with NOAE recorded as $19 - \sum(BinaryEndpoints)$. All analysis was performed using R [13].

2.2. Aggregate Entropy

The traditional Shannon's entropy $H(X)$ [14], in nat units, is:

Let X be a discrete random variable with a possible set of realizations x, thus;

$$H(X) = -\sum_{x} p(x)\log_e p(x)$$

We define a random variable and its realizations as follows:

For each chemical C at a given concentration, let X_i represent embryo i with $i = 1, \ldots, 32$ and B_j represent biological state j with$j = 1, \ldots, 19$. In addition, X_i has realization x_{ij} with its sample value shown in Fig. 1. The probability mass function can be written as:

$$p\left(B_j|C, X_i\right) = \frac{x_{ij}}{19}$$

The Aggregate Entropy (AggE) for chemical C at a given concentration is summarizing the Shannon's entropy of all tested embryos, which is:

$$AggE = -\sum_{i=1}^{32}\sum_{j=1}^{19} p\left(B_j|C, X_i\right)\log_e\left\{p\left(B_j|C, X_i\right)\right\}$$

2.3. Threshold determination

We first used a chi square approximation to the distribution of AggE of each concentration as well as the distribution of the pooled concentration [15,16]. We estimated our chi square degree of freedom by using the Newton algorithm to optimize the logarithm of the full likelihood of a chi square probability density function. Let $(AggE_1, AggE_2, \ldots, AggE_N)$ be a set of AggE, thus the full likelihood can be written as:

$$f(AggE_1, AggE_2, \ldots, AggE_N) = \left(\frac{1}{2^{\frac{k}{2}}\Gamma(k)}\right)^n \times (AggE_1 * \ldots * AggE_N)^{\frac{k}{2}} e^{-\frac{AggE_1+\ldots+AggE_N}{2}}$$

where k is the degree of freedom of a Chi-square distribution and N is the number of chemicals. Since the maximum likelihood estimator is nonlinear, we first took the negative logarithm of the full likelihood. After that, given a start value, we used Newton iteration to optimize the negative logarithm of the full likelihood such that it gave us the optimal estimate of the degree of freedom of our chi square distribution. Our threshold, which depends on the observed incidences of multiple measurements over many individuals, is the critical value of a one-sided chi square test with the significance level of 0.05.

2.4. Endpoint clustering and sensitivity analysis

We next used pairwise mutual information to characterize the relationship among endpoints. Let E_1 andE_2 represent two endpoints with realization $e_1$and $e_2$as observed incidence counts per chemical-per concentration, given the Shannon's entropy defined above, the joint Shannon's entropy forE_1 and E_2 is:

$$H(E_1, E_2) = -\sum_{e_1}\sum_{e_2} p(e_1, e_2)\log_e p(e_1, e_2)$$

And the conditional entropy can be written as:

$$H(E_1|E_2) = -\sum_{e_1}\sum_{e_2} p(e_1, e_2)\log_e p(e_1|e_2)$$

With all these definitions, the mutual information (MI) is:

$$MI(E_1, E_2) = \sum_{e_1}\sum_{e_2} p(e_1, e_2)\log_e \frac{p(e_1, e_2)}{p(e_1)p(e_2)} = H(E_1) - H(E_1|E_2)$$

MI has the following, commutative, property:

$$MI(E_1, E_2) = MI(E_2, E_1)$$

We formed our clusters based on a modified three-step measurement [17]. First, the pairwise mutual information between endpoints, $MI\left(E_i, E_j\right), i, j = 1, \ldots, 18$, is calculated by using R package "infotheo" [18]. Next, the mutual information matrix is transferred to a distance measurement, called normalized information distance [19], which is:

$$d\left(E_i, E_j\right) = 1 - \frac{MI\left(E_i, E_j\right)}{H(E_i) + H\left(E_j\right) + MI\left(E_i, E_j\right)}$$

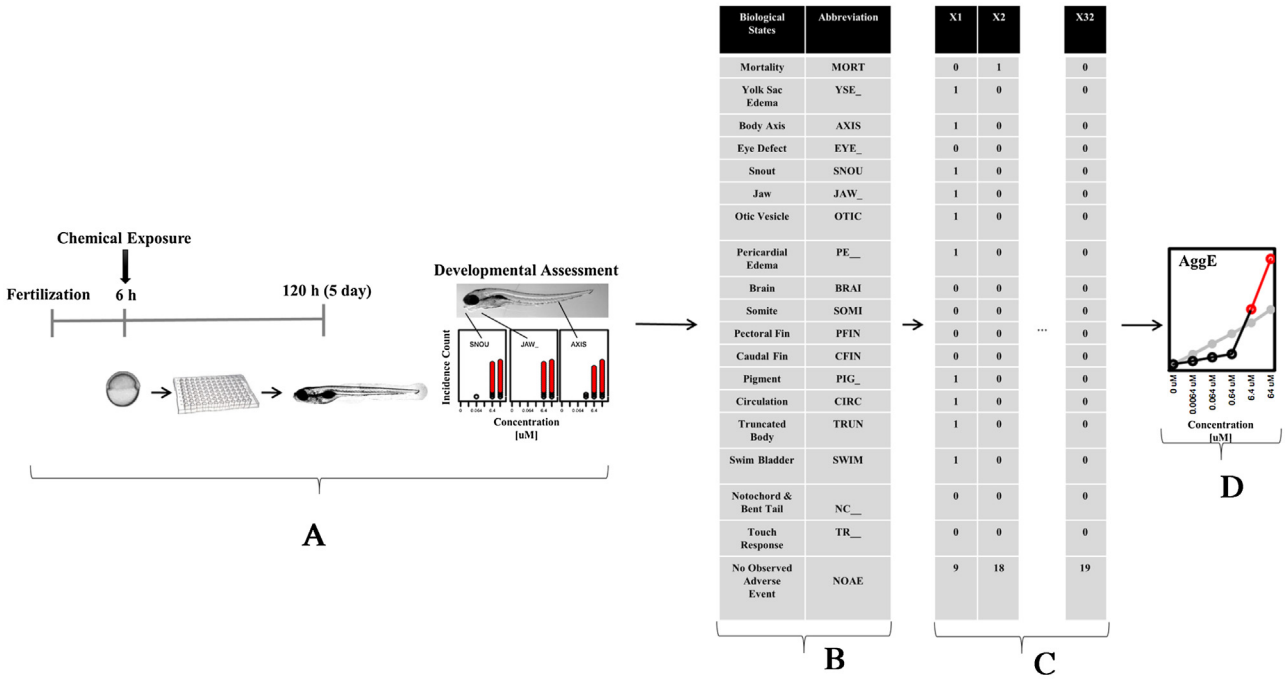

Fig. 1. Experimental Design and Data Structure. (A) Chemical exposure started at 6hpf. At 120 hpf, 18 distinct developmental assessments were measured. (B) 19 biological states (developmental assessments plus NOAE) with their abbreviations. (C) Data structure showing three example vectors from n = 32 individual wells per concentration-by-chemical. X1 indicates many developmental problems observed; X2 shows mortality; X32 represents no phenotypic consequences recorded. (D) Aggregate Entropy, in nats (natural unit of information) on the vertical axis by concentration on the horizontal axis. The black lines connect the concentration-wise AggE for this example chemical, turning red at the point-of-departure concentration, where the line crosses the grey significance threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Finally, hierarchical clustering with normalized information distance and Ward's method was used to characterize the relationship between endpoints.

Our sensitivity analysis followed a three-step procedure. First, based on our clustering analysis, we decided which endpoint or endpoints (super endpoints) we wanted to remove or collapse. Second, we recalculated AggE based on the new set of the endpoints and determined our threshold following the same algorithm defined above. Third, we calculated the concordance (in at least one concentration) between AggE and Fisher's Exact Test for identifying developmental effects.

2.5. Simulation

Given the data structure in Fig. 1, for each endpoint, we simulated a series of Bernoulli trials with sample sizes of 8, 16, 32, 64 and 96 per chemical-per concentration with the real frequencies defined as

$$p(x=1)=\frac{\sum_{i=1}^{32}x_i}{32} and p(x=0)=1-p(x=1)$$

where x_i is the binary incidence for embryo *i* for the given endpoint.

3. Results

As detailed in *Methods*, AggE was developed using data collected according to the experimental design presented in Fig. 1, where each of the 1060 unique chemicals were tested at six concentrations, with n = 32 replicates (wells each containing an individual embryo) at each concentration (Truong et al. [5]). Chemical exposure began at 6 h post fertilization (hpf), then all replicates were evaluated for a suite of 18 developmental endpoints at 120 hpf.

3.1. Distribution and threshold across concentration for AggE

The histograms of AggE across concentrations are shown in Fig. 2. Our chi-square approximation is consistent with the kernel density estimate. The distribution shifts to the right as the concentration increases because we generally observed higher incidence rates at higher concentrations. According to our threshold for AggE (*versus* the univariate Fisher's Exact Test), we found that 24 (*versus* 10) chemicals significantly affected the development of zebrafish at a concentration of 0.0064 μM; 25 (*versus* 15) chemicals at 0.064 μM; 49 (*versus* 34) chemicals at 0.64 μM; 56 (*versus* 59) chemicals at 6.4 μM and 139 (*versus* 168) chemicals at 64 μM. The consequences of mortality are evident in the differences between the distributions at 64 μM (highest observed mortality) and the 'Global' threshold, which was less sensitive to suppression of AggE from observed mortality (see discussion of Fig. 3, below). Table 1 contains information on thresholds and summary statistics.

3.2. Evaluation of AggE in predicting individual morphological effects

For each concentration, we also estimated the general agreement between AggE and Fisher's Exact Test on specific endpoints. We did not include mortality in Fisher's Exact Test, because if an embryo was dead, we could not measure any other endpoints, and our method was designed to evaluate the hazard information across endpoints. For our calculations, *PP* represents tested positive in both tests; *NN* represents tested negative in both tests; *NP* represents tested negative in AggE and positive in Fisher's Exact Test and *PN* represents the opposite case. The balanced ROC (Receiver Operating Characteristic) curve, balanced F1 score measurement $\frac{2\times PP}{2\times PP+PN+NP}$ and concordance $\frac{(PP+NN)}{(PP+NN+PN+NP)}$ between the two tests are shown in Table 1. As an overall summary, the 1060 chemicals are

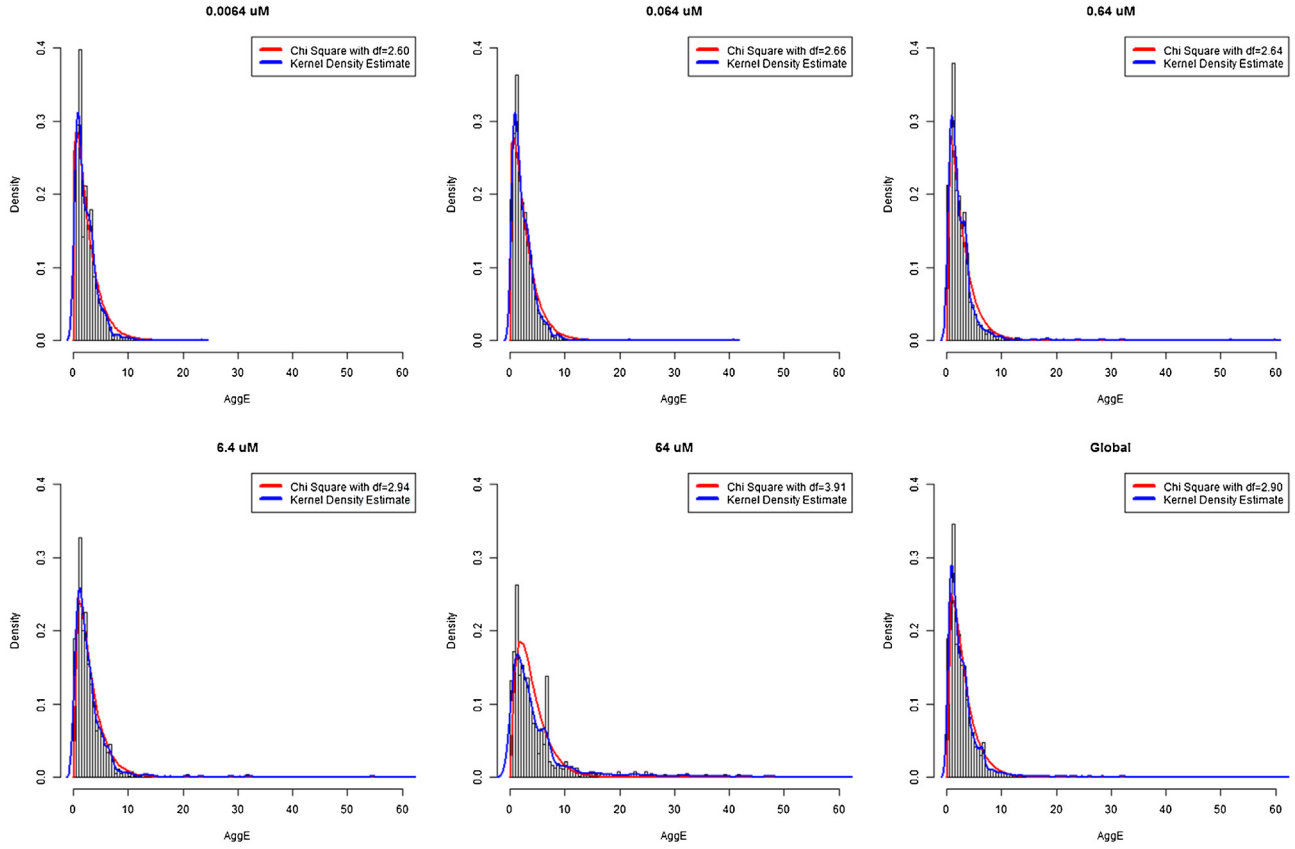

Fig. 2. Histogram of AggE across concentrations. The horizontal axis is AggE, and the vertical axis is the density. The blue line is a kernel density estimate, and the red line is a chi square approximation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Threshold determination of AggE with multiple evaluations including balanced ROC curve, balanced F1 score and concordance.

Concentration	Degree of Freedom (Chi-square)	Threshold (Q(x > 0.05))	# of Significant Chemicals AggE (Univariate Test)	Balanced ROC	Balanced F1 Score	Concordance
0.0064 μM	2.60	7.10	24 (10)	0.56	0.20	0.98
0.064 μM	2.66	7.21	25 (15)	0.70	0.51	0.98
0.64 μM	2.64	7.18	49 (34)	0.73	0.60	0.97
6.4 μM	2.94	7.71	56 (59)	0.88	0.81	0.97
64 μM	3.91	9.35	139 (168)	0.89	0.81	0.95
Global	2.90	7.64				

displayed in decreasing order of their maximum-normalized AggE score, summed across all concentrations Supplementary Table S1 in the online version at DOI: 10.1016/j.reprotox.2016.04.012 (.csv).

3.3. Comparison of AggE with Fisher's exact test

We next observed a positive relationship between the numbers of significant endpoints of each chemical identified by Fisher's Exact Test and its associated Aggregate Entropy (Fig. 3A). We found that AggE is less likely to detect chemicals that cause only mortality, which is expected, given that mortality overwrites all specific endpoints as zero (see Fig. 1). 12-Benzenedicarboxaldehyde (6.4 μM; 64 μM) is shown as an example of this particular case (Fig. 3B), where no concentration-response is evident in the specific endpoints, and only mortality is observed at higher concentrations. When compared to Fisher's Exact Test for each specific endpoint, our method is less likely to detect chemicals where the incidence rate of that endpoint just reaches the significance threshold, as with 5-[2-methyl-3-(pyridine-3-yl)-1H-indol-1-yl]pentanoic acid (Fig. 3B). On the contrary, chemicals having moderate incidence across several endpoints, yet fail to reach the statistical threshold for any single endpoint are identified by AggE. These chemicals have moderate incidence rates across multiple test endpoints and disproportionately affect certain individuals in the population, possibly reflecting genetic variability or experimental difficulty in pathological annotation of several related endpoints. For example, many embryos exhibited developmental endpoints when exposed to Di(2-ethylhexyl) adipate (Fig. 3B); however, none of these incidence rates were significant according to univariate criteria, while from an integration perspective, such a profile warrants concern. We also constructed a new endpoint named "Any_End" to contrast with AggE. "Any_End" represents an observable positive response in any of the tested endpoints and should thus behave similarly to the most sensitive specific endpoint. The Ziram example in Fig. 3B shows AggE accretion over concentrations displaying several specific endpoint responses.

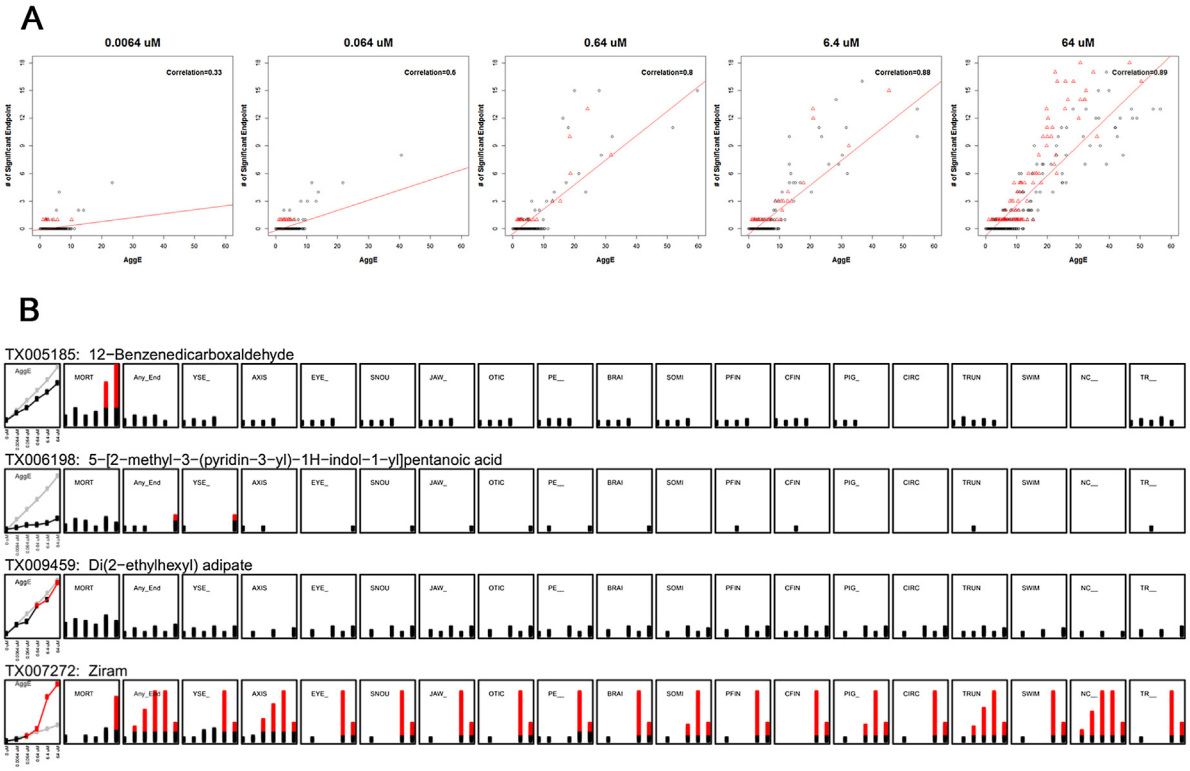

Fig. 3. (A) Correlation between the number of chemicals associated with significant endpoints determined by Fisher's Exact Test (vertical axis) by AggE (horizontal axis). From left to the right, the plots show results for concentrations 0.0064 μM, 0.064 μM, 0.64 μM, 6.4 μM and 64 μM, respectively. Red triangle: significant mortality and/or other specific endpoint(s); black dot: significant endpoint(s) only (except mortality); red line: linear regression fit. (B) similarities and dissimilarities between our method and Fisher's Exact Test on each individual endpoint. For the first panel of each chemical (AggE): gray line is the cumulative summation of the threshold of AggE by concentration; black line is the cumulative summation of chemical associated AggE by concentration, with points colored red that exceed the threshold. For other panels of each chemical: the dot is incidence counts and for a given concentration, if the count is significant by Fisher's Exact Test, it turns red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Clustering analysis

As in Truong et al. [5] a pairwise correlation matrix of the endpoints based on the lowest effect level shows an irregular correlation surface. Here we use an information theory-based approach to identify clusters of the endpoints in order to appropriate handle correlation stemming from individual zebrafish profiles as well as endpoint relatedness. The pairwise mutual information matrix is shown in Supplementary Table S3 in the online version at DOI: 10.1016/j.reprotox.2016.04.012. We next followed the procedures described in the methods section to find clusters with similar phenotypic responses (Fig. 4A). From both the mutual information across endpoints and clustering analysis, notochord distortion (NC), bent body axis (AXIS), touch response (TR), and mortality (MORT) seem to be independent of other endpoints. The other 5 clusters include craniofacial endpoints (Eye, Snout and Jaw), edemas (Yolk Sac Edema and Pericardial Edema), upright body (Swim Bladder, Somite and Circulation), Brain (Brain, Otic Vesicle and Pectoral Fin), and Trunk (Trunk and Caudal Fin).

We performed a sensitivity analysis by removing one endpoint at a time or a cluster of endpoints (SE: Super Endpoints). After removing one endpoint at a time, we found that we do not lose the power of detecting that particularly removed endpoint due to the high mutual information shared with any other endpoint(s). However, this is not true for removing a single clustered endpoint. For example, the mutual information for two edema endpoints is very high. After removing these two endpoints, we found that we lost the power of detecting chemicals previously associated with edema. However, we increase the power of detecting chemicals that caused other developmental defects because of the reduced noise of the data caused by the irregular correlation structure. This trend continues after the same analysis over other super endpoints. Based on this fact, we carried out an analysis on our 10 super endpoints (Fig. 4A), which are the 10 clusters defined above. For any super endpoint that contains more than one single endpoint, if at least one developmental defect was observed within the same super endpoint of that embryo, we recorded that this embryo has this particular defect. For instance, Edema contains two single endpoints (YSE and PE). If one embryo has either one or both, we state that this embryo has an edema problem. We compared the balanced sensitivity, specificity and F1 score of our method on the new super endpoints with the original single endpoint in classifying chemicals that have a significant effect on a specific endpoint based on Fisher's Exact Test (Fig. 4B). In general, our method performed better using super endpoints on any measurement and retained high power for detection of hazardous chemicals. Since mortality supersedes recording of specific developmental endpoints, we also performed the same analysis after removing mortality, resulting in 17 single endpoints and 9 super endpoints. In brief, AggE still performed better using super endpoints, and we increased the power of detecting hazardous chemicals that caused significant developmental problems (Fig. 4B). This reflects the flexibility of AggE using reduced endpoint sets, or more general annotation of difficult-to-discern specific endpoints (*i.e.* annotation as "Edema" *versus* separate YSE or PE entries).

3.5. Simulation

We explored the applicability of our method to different experimental designs by generating simulated data sets with different

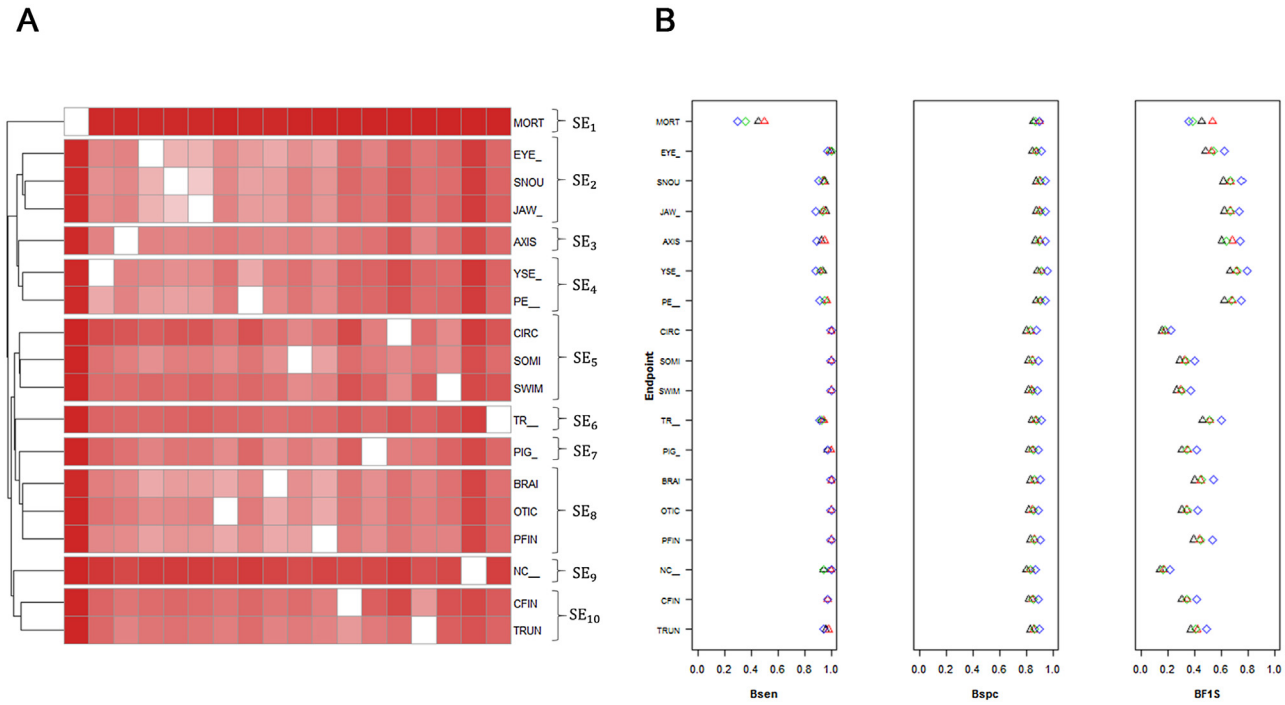

Fig. 4. (A) Heatmap showing hierarchical clustering using normalized information distance with ward linkage. SE: Super Endpoint. (B) Comparison of the predictive power of chemicals that caused significant morphological effect by applying our method on single endpoint with mortality (black triangle); without mortality (green diamond) *vs.* super endpoint with mortality (red triangle); without mortality (blue diamond). Note that only the super endpoint (red triangle, blue diamond) will be visible for cases of perfect overlap with single endpoints. Bsen: Balanced Sensitivity; Bspc: Balanced Specificity; BF1S: Balanced F1 Score. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sample sizes of n = 8, 16, 32, 64 and 96. We compared the variation of our simulated AggEs over the pooled concentration by using violin plots (Fig. 5 Fig. 5). AggE relies on the tested sample size as well as observed incidence rates of multiple measurements. Thus, the degrees of freedom of our chi square approximation to AggE increases with the sample size, because as we attain more embryos, we increase our hazard information (Supplementary Table S2 in the online version at DOI: 10.1016/j.reprotox.2016.04.012). The three measurements, which are balanced ROC curve, F1 score and concordance, all reach to a comparable stationary phase at the sample size of 32 and 64. If the sample size gets too small, balanced ROC curve and concordance become over-representative. Once the sample size gets too big, all three measurements decrease dramatically compared to the raw data measurements in Table 1. In general, we need a big sample size to reduce the bias and get a more accurate estimate. However, in this case, if the sample size gets too big, even a small difference of incidence rate between two experiments can be significant using Fisher's Exact Test, binomial test, or other uncorrected, univariate tests. Thus, these statistical tests on each specific endpoint may not be appropriate, while AggE is still valid regardless of the sample size and can be appropriately parameterized by sample size and observed incidences on multiple measurements.

3.6. Validation

We next tested our method based on the results of an external dataset of flame retardant chemicals [9]. The data structure is comparable to what we have shown here, with n = 32 and the same set of endpoints. For chemicals that have the same tested concentrations as ours, we used the same threshold concentration-wise in Table 1. For those chemicals that have different tested concentrations (6.4E−6 μM, 6.4E−5 μM, 6.4E−4 μM), we used the global threshold in Table 1. The analysis was redone using Fisher's Exact Test, and the chemicals showing significant morphological effects associated with their effective concentrations are displayed in Table 2. The results also show a strong agreement between our method and Fisher's Exact Test. Our method identified three new chemicals (5-OH-BDE-47, BDE100, and TBP) showing evidence of aggregate developmental hazard. There are three chemicals, DE71 (a mixture of brominated diphenyl ethers) at 64 μM, o-TCP (Tri-o-cresyl phosphate) at 64 μM and TDCPP (Tris(1,3-dichloro-2-propyl) phosphate) at 64 μM, that did not reach our AggE threshold but are significant according to Fisher's Exact Test. o-TCP and TDCPP have a very high mortality rate at 64 μM, and DE71 only univariately significant on body bent axis.

4. Discussion

We presented a scoring framework called Aggregate Entropy to evaluate the developmental toxicity of chemicals *in vivo*. In terms of sensitivity, AggE is consistent with Fisher's Exact Test and other contingency-table methods in many scenarios but has advantages when presented with interindividual heterogeneity and endpoint–endpoint correlation. In terms of specificity, AggE reduces potentially false-positive significance calls arising from small numbers in any one cell of a contingency table, rendering AggE more stable in the face of smaller sample sizes and single-endpoints. AggE considers the information of all phenotypic responses of zebrafish after chemical exposure. This aligns with the logic that if a chemical elicits responses in many of the tested endpoints, yet none of these singular endpoints reaches the incidence threshold by Fisher's Exact Test, we should still annotate its potential hazard. Due to our limited knowledge about the underlying biological processes perturbed by most chemicals, and because many of the endpoints share elements of the same Adverse Outcome Pathways (AOPs), these chemicals warrant further scrutiny.

Our clustering analysis and sensitivity analysis indicated that there is a strong, yet not uniform, relationship among many

Table 2
Validation of our method using Noyes et al. [9].

Chemical Name	Concentration(μM)	Significant Endpoint(s)	Aggregate Entropy
BPDP	64	YSE, AXIS, EYE, SNOU, JAW, PE, PFIN, CFIN	37.91[c]
mITP	0.64	YSE, AXIS, SNOU, JAW, PE, PFIN, SWIM, TR	31.78[c]
IPP-1	64	MORT, YSE, AXIS, PE, PFIN, CFIN, TR	23.32[c]
IPP-3	64	YSE, AXIS, PE, PFIN, CFIN	20.41[c]
TBBPA	6.4	MORT, AXIS, JAW, CFIN, TRUN, TR	12.23[c]
IPP-2	64	YSE, AXIS, PE	10.94[b]
TCP	64	MORT, YSE, AXIS, PE, TR	10.94[b]
5-OH-BDE-47	0.00064	None	7.28[d]
BDE100	0.00064	None	6.43[d]
TBBPA	64	MORT	6.39[d]
TDCPP	64	MORT, CFIN	5.92
o-TCP	64	MORT	5.73
TBP	0.0064	None	5.66[a]
DE 71	64	AXIS	3.3

[a] Significant at 0.1.
[b] Significant at 0.05.
[c] Significant at 0.001.
[d] Significant at 0.1 using Global Threshold (Concentration does not match).

endpoints in these data, which is especially common in developmental studies where a coordinated cascade of biological events must take place. We need to have methods that do not inflate false-positives nor lose power (*i.e.* inflate false-negatives) when faced with irregular correlation structure. AggE was designed to solve this problem. Across a diverse chemical set, we can capitalize on this correlation structure to hypothesize endpoints related by common perturbations or adverse outcome pathways. In addition, we showed the benefits of removing specific endpoints that shared an especially tight correlation structure with other endpoint(s). The analysis on 10 super endpoints outperformed (measured by detection of previously-identified chemical effects) the results using the full set of original, specific endpoints. This may aid future experimental design by negating the need to annotate difficult-to-separate endpoints into specific bins or enable implementation of fully-automated annotation protocols.

Our method offers several benefits over common statistical methods used in analyzing zebrafish morphological data. First, we were able to detect chemicals having robust effects on specific endpoints based on Fisher's Exact Test, as well as many new chemicals that would be missed by such traditional methods. Second, AggE maintains appropriate detection power when faced with extremely large or small sample sizes, whereas contingency-table methods suffer an inflation of false-positives. Third, AggE does not enforce a global model, whereas simple linear or logistic fit models will not be appropriate in data where the variance of the incidence rate is not constant and residuals differ across concentrations. Fourth, if we simply add all of the observed incidences for each embryo then perform a standard statistical test on the summation, the results can be misleading due to the fact that the same event will be over-counted because of high shared mutual information across endpoints. This is a salient feature of developmental assays, where some key event(s) can trigger many observable phenotypes. Fifth, AggE can be applied to datasets of varying size, complexity, and degree of non-independence, since its threshold is a function of observed incidences over many individuals. We have demonstrated its use in a high-dimensional zebrafish development assay, but this method could be applied to multiplexed measurements in other *in vitro* or *in vivo* systems, or even to binarized "hits" from assay suites, gene expression, or pathway enrichment analysis.

5. Conclusions

In summary, we developed a new computational approach to characterize chemical exposure information and applied it scoring multiplexed, phenotypic endpoints measured in zebrafish (*Danio rerio*) across several concentrations. We were able to elucidate multi-endpoint syndromes across related endpoints as well as identify chemicals that displayed generalized teratogenic effects. As a complement to rank-based [10], curve-fitting [11], and *a priori* weighting metrics [12], AggE is a flexible approach that is capable of identifying hazardous chemicals from data encompassing a broad parameter space, while avoiding many statistical pitfalls of traditional methods. By testing AggE in a variety of high-dimensional real and simulated datasets, we have characterized its performance and suggested implementation parameters that can guide its application across a wide range of experimental scenarios.

Conflicts of interest

The authors declare that they have no conflict of interest.

Transparency document

The Transparency document associated with this article can be found in the online version.

The Transparency document associated with this article can be found in the online version.

Acknowledgements

This work was supported by NIEHS grants R01 ES19604, R01 ES023788, P42 ES005948, P30 ES025128, RC4 ES019764 P30, P30 ES000210, P42 ES016465, 5T32ES007329, and Environmental Protection Agency (EPA) STAR Grants #835168 and #83579601.

References

[1] B.J. George, D.M. Reif, J.E. Gallagher, C.R. Williams-DeVane, B.L. Heidenfelder, E.E. Hudgens, W. Jones, L. Neas, E.A. Hubal, S.W. Edwards, Data-Driven asthma endotypes defined from blood biomarker and gene expression data, PLoS One 10 (2) (2015) e0117445, http://dx.doi.org/10.1371/journal.pone.0117445.
[2] F.S. Collins, G.M. Gray, J.R. Bucher, Toxicology. Transforming environmental health protection, Science 319 (5865) (2008) 906–907, http://dx.doi.org/10.1126/science.1154619.
[3] R.S. Judson, K.A. Houck, R.J. Kavlock, T.B. Knudsen, M.T. Martin, H.M. Mortensen, D.M. Reif, D.M. Rotroff, I. Shah, A.M. Richard, D.J. Dix, In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project, Environ. Health Perspect. 118 (4) (2010) 485–492, http://dx.doi.org/10.1289/ehp.0901392.
[4] D.M. Reif, L. Truong, D. Mandrell, S. Marvel, G. Zhang, R.L. Tanguay, High-throughput characterization of chemical-associated embryonic behavioral changes predicts teratogenic outcomes, Arch. Toxicol. (2015), http://dx.doi.org/10.1007/s00204-015-1554-1.

[5] L. Truong, D.M. Reif, L. St Mary, M.C. Geier, H.D. Truong, R.L. Tanguay, Multidimensional In vivo hazard assessment using zebrafish, Toxicol. Sci. 137 (1) (2014) 212–233, http://dx.doi.org/10.1093/toxsci/kft235.
[6] G.J. Lieschke, P.D. Currie, Animal models of human disease: zebrafish swim into view, Nat. Rev. Genet. 8 (5) (2007) 353–367.
[7] K. Howe, et al., The zebrafish reference genome sequence and its relationship to the human genome, Nature 496 (2013) 498–503, http://dx.doi.org/10.1038/nature12111.
[8] A.J. Rennekamp, R.T. Peterson, 15 years of zebrafish chemical screening, Curr. Opin. Chem. Biol. 24 (2015) 58–70, http://dx.doi.org/10.1016/j.cbpa.2014.10.025.
[9] P.D. Noyes, D.E. Haggard, G.D. Gonnerman, R.L. Tanguay, Advanced morphological-Behavioral test platform reveals neurodevelopmental defects in embryonic zebrafish exposed to comprehensive suite of halogenated and organophosphate flame retardants, Toxicol. Sci. 145 (1) (2015) 177–195, http://dx.doi.org/10.1093/toxsci/kfv044.
[10] D.M. Reif, M.T. Martin, S.W. Tan, K.A. Houck, R.S. Judson, A.M. Richard, T.B. Knudsen, D.J. Dix, R.J. Kavlock, Endocrine profiling and prioritization of environmental chemicals using ToxCast data, Environ. Health Perspect. 118 (12) (2010) 1714–1720, http://dx.doi.org/10.1289/ehp.1002180.
[11] S. Padilla, D. Corum, B. Padnos, D.L. Hunter, A.L. Beam, K.A. Houck, N. Sipes, N. Kleinstreuer, T. Knudsen, D.J. Dix, D.M. Reif, Zebrafish developmental screening of the phase I chemical library, Reprod. Toxicol. 33 (2) (2012) 174–187, http://dx.doi.org/10.1016/j.reprotox.2011.10.018.
[12] B. Harper, D. Thomas, S. Chikkagoudar, N. Baker, K. Tang, A. Heredia-Langner, R. Lins, S. Harper, Comparative hazard analysis and toxicological modeling of diverse nanomaterials using the embryonic zebrafish (EZ) metric of toxcity, J. Nanopart. Res. 17 (6) (2005) 250, http://dx.doi.org/10.1007/s11051-015-3051-0.
[13] R Core Team, R. A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015, URL http://www.R-project.org/.
[14] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[15] B. Goebel, Z. Dawy, J. Hagenauer, J.C. Mueller, An approximation to the distribution of finite sample size mutual information estimates, ICC 2 (2005) 1102–1106, http://dx.doi.org/10.1109/ICC.2005.1494518.
[16] V.P. Singh, Entropy Theory and Its Application in Environmental and Water Engineering, John Wiley, New York, 2013.
[17] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, J.C. Mueller, Gene mapping and marker clustering using shannon's mutual information, IEEE/ACM Trans. Comput. Biol. Bioinform. 3 (January–March (1)) (2006) 47–56.
[18] P.E. Meyer, infotheo: Information-Theoretic Measures, R package version 1.2.0., 2014, http://CRAN.R-project.org/package=infotheo.
[19] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Viranyi, The similarity metric, IEEE Trans. Inf. Theory 50 (12) (2004) 3250–3264.